Distinction of 3D Objects and Scenes via Classification Network and Markov Random Field

Ran Song, Yonghuai Liu, *Senior Member, IEEE* and Paul L. Rosin

Abstract—An importance measure of 3D objects inspired by human perception has a range of applications since people want computers to behave like humans in many tasks. This paper revisits a well-defined measure, distinction of 3D surface mesh, which indicates how important a region of a mesh is with respect to classification. We develop a method to compute it based on a classification network and a Markov Random Field (MRF). The classification network learns view-based distinction by handling multiple views of a 3D object. Using a classification network has an advantage of avoiding the training data problem which has become a major obstacle of applying deep learning to 3D object understanding tasks. The MRF estimates the parameters of a linear model for combining the view-based distinction maps. The experiments using several publicly accessible datasets show that the distinctive regions detected by our method are not just significantly different from those detected by methods based on handcrafted features, but more consistent with human perception. We also compare it with other perceptual measures and quantitatively evaluate its performance in the context of two applications. Furthermore, due to the view-based nature of our method, we are able to easily extend mesh distinction to 3D scenes containing multiple objects.

Index Terms—3D mesh, distinction, neural network, Markov Random Field.

1 Introduction

Measures of regional importance of 3D surfaces in agreement with human visual perception have been an active research topic in computer vision and graphics for quite a long time. Popular measures include mesh saliency [1], mesh distinction [2], 3D interest points [3], Schelling distribution [4], etc. These measures represent understanding of 3D objects from different perspectives and have led to a range of applications such as best view selection [5], [6], mesh segmentation [7], mesh simplification [1], [2], scan integration [8] and 3D printing [9].

This paper focuses on a measure of mesh distinction. As defined in [2], it measures how important a mesh region is for distinguishing a 3D object from others of different classes. In [2], the distinction is computed through handcrafted local features while in our method, it is learned through a Convolutional Neural Network (CNN) for object classification. This is motivated by two facts. First, perceptual measures are usually influenced by both local and global features. CNN-based methods have demonstrated that they are good at extracting such features and making good balance between them in the tasks of computing perceptual measures for images [10], [11]. Second, handcrafted features normally do not generalise well since their expressivity capabilities are limited by the fixed operations that stay the same for different sources of data. In contrast, CNNs learn features specific to the data and have demonstrated strong generalisation capability through various applications. It is worth noting that usually the CNN has to be sufficiently deep to generalise well. This is because only a sufficiently large number of layers can learn features at various levels of abstraction. However, we noticed that many 'toy' CNNs have been used in computer graphics [12], [13]. They are typically composed of a very small number of layers, which are insufficient to learn all the intermediate features between the raw data and the high-level understanding tasks. Hence, to capture both local and global attributes related to the perceptual distinction measure and guarantee a powerful generalisation capability, we employ a sufficiently deep CNN in this work.

However, without a large training dataset, training a deep CNN will lead to overfitting which will weaken its generalisation capability. Unfortunately, acquiring a large training dataset is usually very difficult in 3D object understanding tasks [4], [6], [13]. But object classification is an exception as acquiring a large training dataset for it is practically easy. Moreover, compared with human-generated vertex-level annotation, object-level annotation (i.e. class membership of 3D object) is more reliable and objective since it is almost free from inconsistency. We propose to estimate distinction through classification network, which provides a practical and novel method to address the training data problem widely existing among various 3D object understanding tasks.

Inherently, our method is weakly supervised where we try to estimate vertex-level annotation based on object-level annotation. This idea is not new in image segmentation. Since collecting a large amount of pixel-level annotation is difficult, researchers proposed to use image-level annotation to address semantic segmentation problems [14], [15], [16]. However, our problem is much more challenging. In semantic image segmentation, the pixel-level annotation to be estimated is not just discrete, but also fully consistent with image-level annotation. In contrast, the vertex-level annotation for mesh distinction is not from a set of known

- *R. Song (email: r.song@brighton.ac.uk) is with the Centre for Secure, Intelligent and Usable Systems, School of Computing, Engineering and Mathematics, University of Brighton, UK*
- *Y. Liu is with the Department of Computer Science, Edge Hill University, UK.*
- *P. L. Rosin is with the School of Computer Science and Informatics, Cardiff University, UK.*

Manuscript revised Nov 2018

class names but an unknown real number. Compared to pixel-level annotation, acquiring vertex-level annotation is usually more labour intensive since some vertices are not immediately visible and the human subjects have to rotate and zoom the 3D mesh to view and annotate them. Therefore, although more challenging, a weakly-supervised solution is usually more preferable for 3D object understanding tasks.

Certainly, since distinction and object classification are two different tasks, a fundamental issue is the transferability of the knowledge learned through classification networks, which has only been explored in the context of 2D image understanding [17], [18]. One consensus is that the transferability decreases as the distance between the base task and the target task increases [17]. Hence, our method works well because the knowledge vital for the base task, i.e. classification, is highly consistent with the target task, i.e. distinction.

We propose to estimate mesh distinction through a sufficiently deep classification network trained on a large number of 3D objects. First, inspired by the multi-view CNN [19] which performed well in 3D object classification [20] and semantic segmentation [21], we represent the 3D object as a sufficiently large number of 2D views so that the loss of information caused by the projection from 3D to 2D will not cause a significant sacrifice in representation. Second, each view is fed into a CNN to generate a 2D distinction map through back-propagating the classification probability of each view corresponding to the correct object class. Third, we convert a 2D distinction map into a 3D one by considering the visibility of the vertices with respect to the corresponding viewpoint. Finally, different from multi-view CNN which merely uses a simple max-pooling scheme, we construct a Markov Random Field (MRF) which takes into account the pairwise relations of the views and then infer it to incorporate multiple distinction maps into a single one.

Moreover, we extend distinction to 3D scenes. To the best of our knowledge, all of the previous methods on perceptual measures for 3D meshes only concerned a single object. For the first time, we extend the concept from one single object to a scene which contains multiple objects. It is worth noting that perceptual measures of a 3D scene is not a new topic but researchers have only computed them based on depth images [22], [23], [24]. The disadvantage of depth-based perceptual measure of a 3D scene is that the resultant perceptual map is only valid for a particular view. The proposed 3D scene distinction is valid for all views.

Overall, the contribution of our work is threefold:

1) We propose a novel method for computing mesh distinction and show its applications.
2) We propose a novel multi-view architecture where an MRF is developed to combine information learned across multiple views.
3) We extend the concept of distinction from single object to 3D scenes containing multiple objects and describe a practical method to compute it.

2 Related work

Measures of perceptual importance of 3D surface regions have been inspired by an extensively studied area in computer vision known as image saliency [25], [26], [27]. While image saliency explores colour and temporal coherence, measures for 3D surfaces reason about the geometry of meshes. Thus early works in this field heavily rely on local geometric features. For example, Lee *et al.* [1] computed mesh saliency using a center-surround operator on Gaussian-weighted curvatures calculated in a local neighbourhood at multiple scales, and was later demonstrated in [28] that such a mechanism has significantly better correlation with human eye fixations than either a random model or curvatures. Gal and Cohen-Or [29] introduced a salient geometric feature based on curvatures which functionally characterizes a local partial shape. It was then used to improve part-in-whole shape matching. Shilane and Funkhouser [2] developed an approach to compute the distinctive regions of a 3D surface by describing the shape of every local spherical region through a Harmonic Shape Descriptor. Castellani *et al.* [30] proposed a method for detecting and matching locally salient points from multi-view meshes, where saliency is determined by generating a multi-scale representation of a mesh in a difference of Gaussian scale space.

While local geometric features indeed influence where people focus their attention in an image or a mesh, saliency actually depends on a few basic principles of human visual attention, as shown by psychological evidence [31], [32], [33]. These include not only local but also global considerations. Thus recently, methods integrating global cues have been proposed although such 'global' perspective is still subject to a single object rather than a scene. Leifman *et al.* [6] proposed an algorithm for detecting salient regions and explored how to select viewpoints based on these regions. Their algorithm looks for regions that are distinct both locally and globally where the global cue is if the object is 'limb-like' or not. Wu *et al.* [34] proposed an approach for detecting mesh saliency based on the observation that salient features are both locally prominent and globally rare. Song *et al.* [8] presented a method considering both local geometric cues and global information corresponding to the log-Laplacian spectrum of a mesh. Saliency is then determined by transferring salient information from the spectral domain back to the spatial domain. Wang *et al.* [7] detected mesh saliency using low-rank and sparse analysis in a shape feature space. The shape features encode both the local geometry and the global structural information of a mesh. Song *et al.* [35] proposed a local-to-global scheme to integrate both the local mesh saliency and the global distinctness of features.

In our work, we estimate distinction through a collection of 2D views of a 3D object. The views record information from the entire object which can enable the discrimination of the object from those in other classes. This is fundamentally different from those in the literature that consider local and/or global features over the objects themselves. The distinction is then estimated by linearly combining such information from different views with weights inferred globally through an MRF.

Differences from relevant works. Chen *et al.* [4] proposed a regression model to predict the so-called Schelling distribution learned on a collection of 400 meshes.

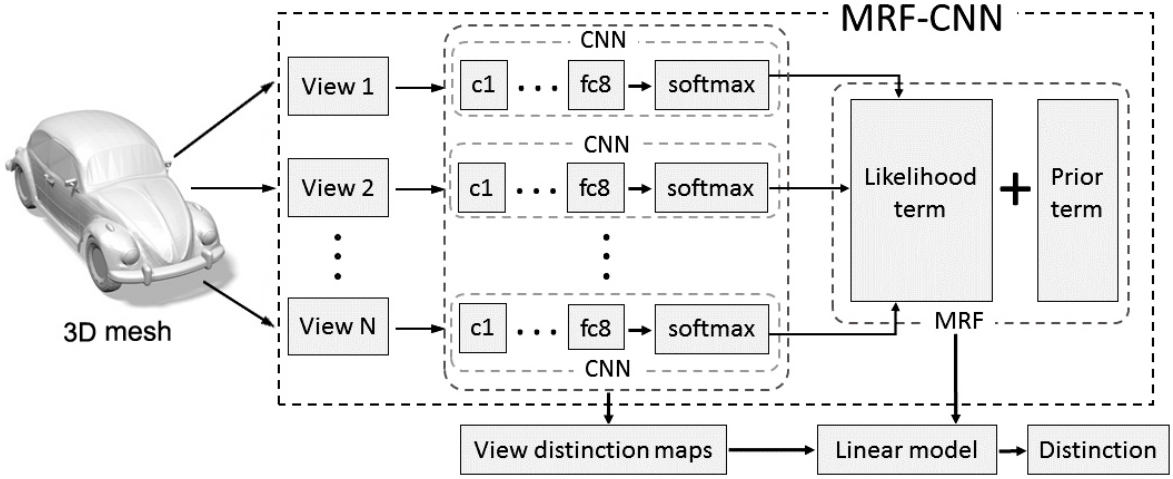


Fig. 1. Overview of the proposed method for estimating mesh distinction.

it is essentially a shallow learning based on a selection of handcrafted features while our method relies on features learned through a deep CNN.

Lau *et al.* [13] proposed the concept of tactile mesh saliency which facilitated a relatively reliable data collection since the concept is well defined and human subjects tend to give highly consistent responses in the process of data collection. Even so, only 150 meshes were collected for both training and testing. As a result, merely a 6-layer neural network was used because such a small number of training data is far from enough to support the learning of a sufficiently deep neural network which contains millions of parameters. Hence, our method is significantly different because our network is much deeper and trained on a much larger dataset, leading to good results on a variety of objects.

Su *et al.* [19] proposed the multi-view CNN for 3D object classification and retrieval. Our method is closely related to it since we also use a similar multi-view setup. The major contribution of Su's work is the proposed 'view-pooling' layer which gathers information learned from multiple views to deliver a single output. Our method is however fundamentally different since we propose an MRF to aggregate such information as illustrated in Fig. 1. To demonstrate the difference between Su's method and ours solidly, we designed an experiment where we replaced the MRF with Su's view-pooling layer and reported the comparative evaluation in Section 4.10. In summary, our method is essentially different from other multi-view CNNs. For instance, Zhu *et al.* [36] simply concatenated the information learned from multiple views for 3D object retrieval. Kalogerakis *et al.* [21] proposed a 'projection layer' to aggregate information across multiple views for 3D shape segmentation. Basically, the projection layer is still a pooling layer. Note that they also proposed a probabilistic model, a Conditional Random Field but it is used for handling information at the vertex level after the aggregation rather than combining knowledge learned from multiple views.

Apart from multi-view CNN there are also other approaches to conducting 3D deep learning. For example, Wu *et al.* [37] proposed a volumetric CNN for object recognition where a 3D object is represented as a volumetric grid. Qi *et al.* [38] developed the PointNet which relies only on the coordinates of vertices for classifying and segmenting 3D objects. However, according to [20] and [38], both of the methods are outperformed by the multi-view CNN of Su *et al.* [19] in classification tasks. Another category of methods is graph neural network (GNN) widely used in geometric deep learning. A GNN learns a deep representation over the meshed surface treated as a non-Euclidean graph by a local operator such as Laplacian [39], [40]. But such local operators are not good at capturing the global positional relationship of multiple objects in a scene. For instance, moving some disconnected objects in a scene will not change the Laplacian of the scene since it is based on the adjacency matrix. In comparison, this relationship is recorded by the 2D views of the scene, which facilitates us to extend mesh distinction to scene distinction. Thus a multi-view setup is particularly suitable for our work since it does not only have a state-of-the-art performance on 3D object classification, but also simplifies the computation of distinction for a scene containing multiple objects.

3 Distinction of 3D meshes via classification network and MRF

We combine a classification network and an MRF to compute a measure which maps from a mesh vertex to a distinction value. The network architecture is the classic VGG-M model [41] which is a 19-layer CNN. We leverage the VGG-M model to generate 2D distinction maps via training on a large 3D classification dataset. These 2D distinction maps encode distinctive information of the 2D views of a 3D mesh. We also provide a method to associate such information with mesh vertices. The MRF combines a likelihood term derived from the classification network and a prior term to estimate a linear model for aggregating view-based distinction maps to output a per-vertex distinction map.

Fig. 1 illustrates our method based on an MRF-CNN architecture and a linear model. We use the output of the CNN to compute the likelihood term of the MRF. Each view distinction map can be regarded as an attribute of the 3D object. Secord *et al.* [5] demonstrated that a simple

linear model works very well in order to combine multiple attributes for predicting view preference of 3D objects. Therefore, we propose to use a linear model estimated by the MRF to incorporate multiple view distinction maps into a single per-vertex distinction map for the input object.

In the following, we shall describe all steps of the proposed method in detail, including the multi-view mesh representation, the calculation of pixel distinction based on back-propagation, the transfer from pixel distinction to vertex distinction and the combination of multiple vertex distinction maps via an MRF.

3.1 Multi-view mesh representation

We start with an icosahedron to uniformly sample a view sphere which surrounds the input 3D mesh. Then we iteratively subdivide the icosahedron to produce more vertices (i.e., viewpoints) on the view sphere. We finally end up with a polyhedron with N ($N = 42$ in this work) vertices which uniformly samples the view sphere. Using more views usually improves the performance, particularly for concave surfaces. But it will also slow down the method. To create a multi-view representation for a 3D mesh, we place a camera at each vertex of the sampled view sphere, pointing towards the centroid of the mesh, and add a light source to generate a rendered 2D view of the mesh with details described below in Section 4.3. Note that using different shading coefficients or illumination models does not affect our method due to the invariance of the learned convolutional filters to illumination changes, as observed in image-based CNNs. Adding more or different viewpoints is trivial, however, we found that such a rendering setup was already enough to achieve high performance.

From Fig. 1, it can be seen that the operation of the CNN on each view is independent. We thus take an MRF into account to formulate the pairwise relations of the views through a prior. In computer vision, an MRF usually models pixel labels as random variables when conditioned upon a global observation. In our problem, the N viewpoints on the sampled view sphere form a graph since they are connected by the edges of the polyhedron.

3.2 View-based pixel distinction via back-propagation

Once we generate multiple views of a 3D mesh, we can estimate the contribution of each view to the correct classification of the mesh. The output of the softmax layer of each CNN is a C-dimensional vector where C denotes the total number of object classes. Each element of the vector represents the probability $P_c(V_n), c = 1, 2...C, n = 1, 2, ...N$ that the 3D object rendered in the view V_n belongs to a certain class c. We obtain a C-dimensional vector by doing an element-wise aggregation for all N such vectors and then find the index corresponding to its largest element as the correct class $\mathcal{C}$ (which actually corresponds to the top-1 predicted class) of the 3D object.

For a particular view V_n, we employ the method proposed in [42] to compute a per-pixel distinction map $I(V_n)$ for all of the pixels in V_n based on their influence on the probability $P_{\mathcal{C}}(V_n)$:

$$I(V_n) = \left.\frac{\partial P_{\mathcal{C}}}{\partial V}\right|_{V_n} \quad (1)$$

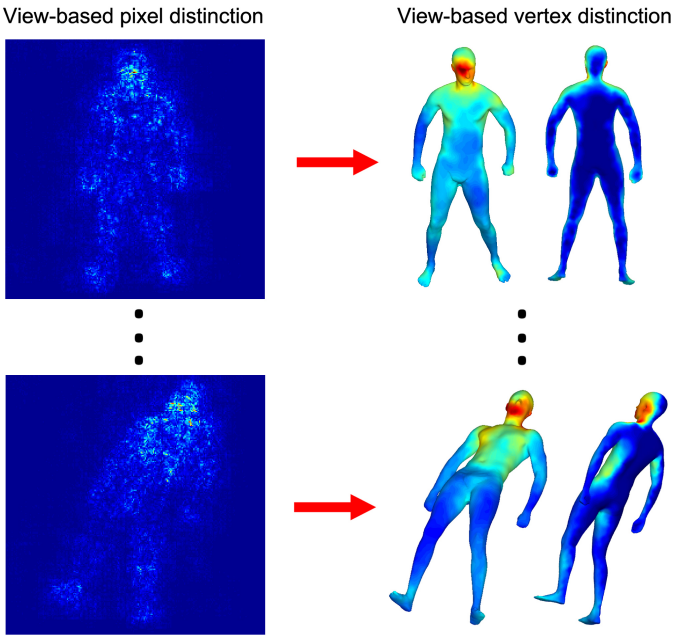


Fig. 2. Transfer from view-based pixel distinction to view-based vertex distinction. Note that the facial region is highlighted in both views and the back of the person is invisible to both viewpoints.

The derivative in Eq. (1) indicates which pixels in V_n need to be changed the least to affect the probability associated with the correct class the most. We found it through a back-propagation with all the CNN parameters fixed after slightly smoothing V_n to remove potential noise using a Gaussian filter with 0 mean, 1.5 standard deviation and a kernel size of 7×7. The distinction map $I(V_n)$ can be interpreted as a measure of pixel importance with regard to the correct classification of the 3D mesh. $I(V_n)$ is further normalised to be within the interval of $[0, 1]$. There is no one-to-one correspondence between the pixels in V_n and the vertices of the 3D mesh. We propose the following method to derive a vertex distinction map from the pixel distinction map for each view.

3.3 Transfer from view-based pixel distinction to view-based vertex distinction

The resolution of the CNN views is fixed (224×224 in this work which empirically shows a good balance between efficiency and performance) no matter how many vertices the 3D mesh contains. Therefore for efficiency, our idea is to compute the view-based vertex distinction maps for a simplified mesh based on the view-based pixel distinction maps. Then we can compute the view-based vertex distinction maps for the original mesh using point correspondence between the simplified mesh and the original mesh.

In detail, we project the simplified mesh at each of the N viewpoints and the 2D projections of all 3D vertices visible to that viewpoint are recorded. Next, we rescale the 2D projections in accordance with the resolution of the CNN views. Then, we crop the rescaled 2D projections to remove the background pixels. Next, the view-based distinction of a 3D vertex is assigned as the view-based distinction of the pixel closest to its 2D projection. Finally, the view-based distinction of a vertex on the original mesh is computed by

finding its closest point on the simplified mesh. As shown in Fig. 2, the view-based pixel distinction maps are typically a little noisy. So before the assignment, we smooth them using a Gaussian filter with 0 mean, 1.5 standard deviation and a kernel size of 7×7. Another way to understand the smoothing is that instead of mapping the distinction of a single pixel to a vertex, we calculate the distinction of a vertex based on pixels within a small neighbourhood. If a vertex is not visible from a viewpoint, its distinction with regard to that view is set to zero. This method results in N view-based vertex distinction maps. A view-based vertex distinction map $D(V_n)$ indicates the importance of each vertex with regard to the correct classification of the object, based on the information recorded in view V_n. For example, in the two views shown in Fig. 2, the vertices in the facial region of the person are of high importance. The back of the person is in dark blue in the two views since it is not visible to the two viewpoints. For simplicity, we write $D(V_n)$ as D_n in the rest of this paper.

Note that the 2D-to-3D transfer of distinction is fundamentally different from the projection layer of the Projective Convolutional Network proposed by Kalogerakis *et al.* [21]. Our method generates a 3D distinction map for each individual 2D distinction map but does not aggregate multiple maps as shown in Fig. 2. Each 3D distinction map is generated based only on the 2D distinction of the corresponding 2D view rather than multiple views and is still view-based. Multiple 3D distinction maps are then aggregated through a newly designed MRF model described in the next subsection. In contrast, the projection layer in [21] aggregates the so-called confidence maps across multiple views via max-pooling. And the output 3D confidence map is not view-based but label-based.

3.4 Combination of multiple distinction maps via MRF

Now we have a collection of N view-based vertex distinction maps. Each distinction map can be interpreted as an attribute which encodes some information of the 3D mesh. Among many potential mathematical models of combining such attributes, an intuitive one is a linear model

$$D = \sum_{n=1}^{N} w_n D_n \tag{2}$$

where w_n denotes the contribution of a view-based vertex distinction map D_n. As a weighting parameter, it reflects the importance of a view in the combination. Secord *et al.* [5] further demonstrated that such a linear model is typically good enough for computing the importance of views for a variety of 3D objects. In their method, the linear model was proposed as a regressor learned on a training dataset since in their work, the likelihoods of the attributes (e.g., projected area, viewpoint entropy, silhouette length, silhouette curvature, depth distribution, etc.) with regard to the importance of views were unknown. However, in our work, we can simply use the class probability output by the CNN as the likelihood (as illustrated in Fig. 1) since it strongly suggests if a view is important or not in terms of distinguishing the object from others of different classes.

As is mentioned above, the output of the final layer, i.e., the softmax layer of the CNN is a vector where each of its elements represents the probability that the 3D object rendered in that view belongs to a certain class. We use the probability corresponding to the top-1 predicted class as the likelihood of the importance of a particular view which reflects how useful it is for distinguishing the 3D object from others of different classes.

Another consideration is that since the views corresponding to neighbouring viewpoints in the MRF graph have similar content, their likelihood of being important or not should also be similar. Hence, we also introduce a prior term to encourage assigning similar weights to neighbouring views in the combination.

We propose a pairwise MRF for the estimation of the weights in Eq. (2) as follows:

$$\begin{aligned} E(\boldsymbol{w}|\boldsymbol{V}) &= \sum_{n=1}^{N} E(w_n|V_n) \\ &= \sum_{n=1}^{N} U_l(V_n|w_n) + \alpha \sum_{n=1}^{N} \sum_{m \in \mathcal{N}(n)} U_p(w_n, w_m) \end{aligned} \tag{3}$$

where $\mathcal{N}(n)$ denotes the neighbourhood of viewpoint n in the MRF graph. α is a parameter which weights the contributions of the likelihood term U_l and the prior term U_p to the MRF energy E. We empirically set it to 0.1 in this work. We formulate the likelihood term as the squared difference between w_n and $P_{\mathcal{C}}(V_n)$ which represents the probability of V_n with respect to the top-1 predicted class $\mathcal{C}$ of the 3D object

$$U_l(V_n|w_n) = (w_n - P_{\mathcal{C}}(V_n))^2. \tag{4}$$

We formulate the prior term as neighbourhood consistency which encourages neighbouring viewpoints to take similar weights and penalises them to take highly different weights

$$U_p(w_n, w_m) = (w_n - w_m)^2 \tag{5}$$

where n and m are neighbouring viewpoints.

Another important benefit of formulating the MRF as the sum of quadratic functions is that it guarantees a linear solution for the maximum likelihood estimate of w_n, which, as we shall show in the next subsection, significantly facilitates the inference of the MRF.

3.5 Inference of the MRF

Our aim is to estimate w_n needed for defining the linear model expressed in Eq. (2). The energy calculated in Eq. (3) is actually the negative logarithm of the posterior probability of the MRF. Maximum a posteriori (MAP) is the most popular principle to infer the MRF in computer vision and graphics, which is equivalent to the minimisation of the energy function. Therefore, the desired configuration of the graph $\widehat{\boldsymbol{w}} = \{\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_N\}$ can be expressed as

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w}) \tag{6}$$

Popular methods for inferring MRFs such as Graph Cut [43] and Belief Propagation [44] require that w_n can only have finite discrete states. However, here w_n are continuous variables. Note that inference in an MRF with continuous states is in general NP hard [45]. But due to the specific formulation of the proposed MRF, we develop a practical

method which usually leads to a solution of a good local optimum although the global optimum cannot be guaranteed.

Since our MRF is pairwise, the conditional distributions of w_n are locally dependent. Thus a desired solution $\widehat{\boldsymbol{w}}$ can be achieved by iteratively assigning each view its local maximum likelihood estimate. This iterative method is known as Iterative Conditional Modes (ICM) algorithm [46]. In our case, the maximum likelihood estimate of w_n can be computed by setting the first derivative of the local MRF equal to zero.

$$\frac{\partial E}{\partial w_n} = \frac{\partial(U_l(V_n|w_n) + \alpha \sum_{m\in\mathcal{N}(n)} U_p(w_n, w_m))}{\partial w_n} = 0. \quad (7)$$

And, because of the quadratic nature of the MRF, its solution is linear, expressed as

$$w_n = \frac{1}{1+\alpha|\mathcal{N}(n)|} P_{\mathcal{C}}(V_n) + \alpha \sum_{m\in\mathcal{N}(n)} w_m \quad (8)$$

where $|\mathcal{N}(n)|$ denotes the number of neighbouring viewpoints of viewpoint n in the MRF graph.

In each iteration, all of the N viewpoints are visited in a random order to avoid propagation of trends. The global root mean square change of w_n is calculated in each iteration and the algorithm terminates when such a change is smaller than a fixed threshold (e.g. 10^{-4}) or the maximum number of iterations $MaxIter$ (e.g. 2000) has been implemented.

It can be seen that an obvious disadvantage of the ICM algorithm is that the inference will be easily trapped at a local optimum. To improve this, we propose two strategies.

First, we modify the ICM algorithm by introducing an annealing strategy. In each iteration, we introduce a small random perturbation to each w_n computed through Eq. (8). Then, if the global MRF energy is reduced, we accept this perturbation; otherwise we compute a temperature (initialised to 1) which is reduced during iterations as

$$T_{Iter} = \frac{T_{Iter-1} \log MaxIter}{\log(MaxIter + Iter)} \quad (9)$$

where $Iter = 1, 2, ..., MaxIter$ and then accept the perturbation with a probability dependent on the temperature: if $P > 1 - T_{Iter}$, update w_n by accepting the perturbation where P is a random value between 0 and 1.

The second strategy is to estimate a good initialisation of w_n. According to Eq. (8), we have

$$\frac{1}{\alpha} w_n - \sum_{m\in\mathcal{N}(n)} w_m = \frac{1}{\alpha + \alpha^2|\mathcal{N}(n)|} P_{\mathcal{C}}(V_n) \quad (10)$$

and benefitting from the specific formulation of the proposed MRF, it can be expressed as a linear system $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$:

$$\begin{bmatrix} \frac{1}{\alpha} & -1 & \dots & -1 & \dots \\ -1 & \frac{1}{\alpha} & -1 & \dots & \\ \vdots & & & & \\ -1 & & & & \\ \vdots & & & & \ddots \end{bmatrix} \boldsymbol{w} = \begin{bmatrix} \frac{1}{\alpha+\alpha^2|\mathcal{N}(1)|} P_{\mathcal{C}}(V_1) \\ \frac{1}{\alpha+\alpha^2|\mathcal{N}(2)|} P_{\mathcal{C}}(V_2) \\ \vdots \\ \frac{1}{\alpha+\alpha^2|\mathcal{N}(N)|} P_{\mathcal{C}}(V_N) \end{bmatrix} \quad (11)$$

where $\boldsymbol{A}$ is a sparse tridiagonal matrix with fringes [47]. Its dimension is $N \times N$ and the positions of -1 can be determined by the adjacency matrix of the polyhedron that forms the MRF graph. Because $\boldsymbol{A}$ is symmetric and positive definite, we apply sparse Cholesky decomposition [48] to solve the linear system and use the solution to initialise the modified ICM algorithm.

4 Experiments, Evaluation and Applications

4.1 Dataset

In order to better evaluate our method, we create a new dataset by expanding the Princeton ModelNet dataset [37]. A 40-class well annotated subset containing 12,311 shapes from 40 common categories, ModelNet40, is publicly available and has been extensively used in 3D object classification and retrieval tasks. In this work, we further extend it to 50 categories by merging it with another popular dataset for perceptual measure of 3D objects. The Schelling distribution dataset [4] contains 400 meshes from 20 categories. 10 of the 20 categories already exist in ModelNet40 and thus we merge these categories from both datasets after aligning the orientations of the objects of each class by the method proposed in [49] and manual effort. For the 10 categories not shared by ModelNet40, we create 10 new categories to accommodate these models in the new dataset after aligning their orientations. We name the new dataset ModelNet50 and training on it enables the evaluation over 20 object classes in Section 4.7.

In the experiments, we use the same training and testing split of ModelNet40 as in [37] where four fifths of the meshes in each category are used for training and one fifth are used for testing. It can be seen that a huge benefit of training deep CNNs based on object classification dataset for computing mesh distinction is that we can easily extend the training dataset (basically by just copy and paste) to make it significantly more diverse, which is usually vital for a generalisation performance of the learning model. In sharp contrast, directly extending other perceptual datasets is very challenging and typically a time-consuming user study with careful consideration about implementation details has to be conducted as suggested in [4]. To further demonstrate the powerful generalisation capability of our method, we also tested it using 3D meshes from other publicly available datasets including the Princeton Shape Benchmark [50], the SHREC'15 (Non-Rigid 3D Shape Retrieval Track) dataset [51], the best view selection benchmark [52] and the 3D interest point detection dataset [3].

4.2 Training strategy

Because we represent a 3D mesh as multiple 2D views, we can leverage massive image databases such as the prevailing ImageNet dataset [53] to pre-train the classification network. Note that the human visual system actually senses 3D geometry through their 2D projections when other necessary elements (e.g., lighting and material) are appropriately provided. And then understanding on 3D geometry can be gained through associating these 2D projections with each other. Hence, on one hand, we can learn a good deal about generic features for 2D image categorisation since images are ubiquitous and large labeled image datasets are abundant; on the other hand, using a deep neural

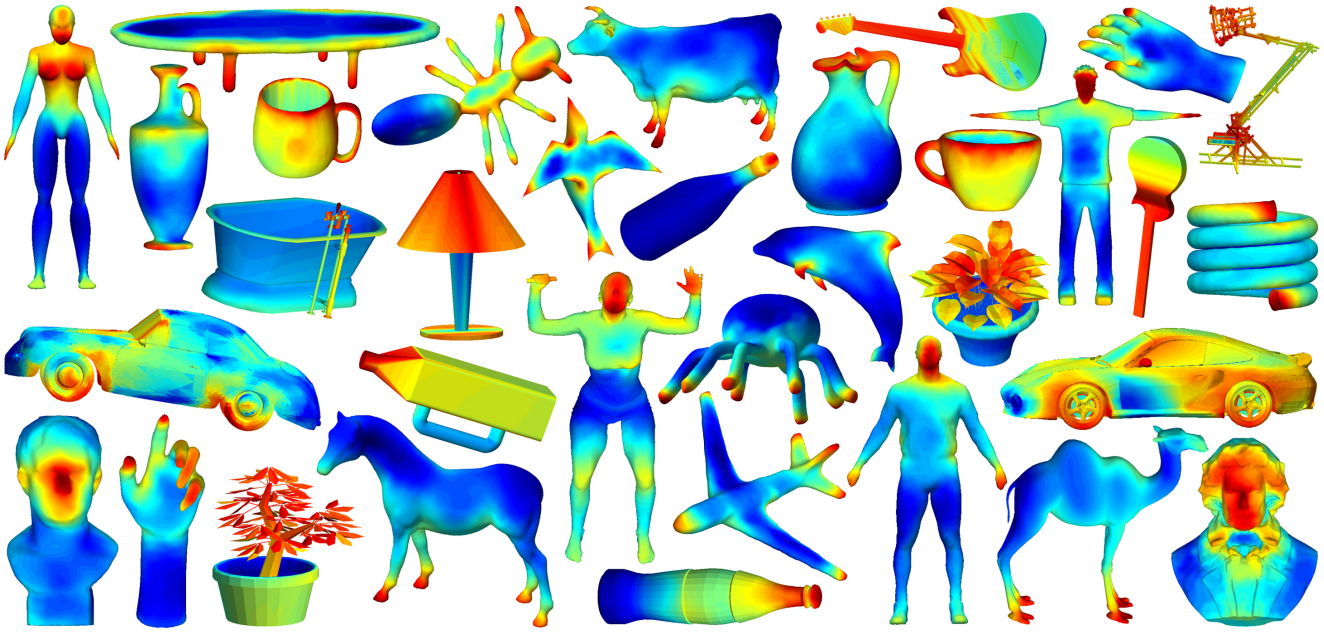


Fig. 3. Distinction maps of various 3D surface meshes in the test dataset of the ModelNet50 dataset

network pre-trained on an image dataset, and then fine-tuning using mesh projections of a dataset of 3D meshes is quite consistent with the mechanism that the human visual system processes visual information. Therefore, we initialise the classification network as the VGG-M [41] network pre-trained on the ImageNet dataset. Then, we fine tune it using the newly created ModelNet50 dataset.

4.3 Rendering details

To render a 3D mesh to a 2D view, we set a camera and a light source at the corresponding viewpoint and use a standard OpenGL renderer with the perspective projection mode. The strengths of the ambient light and the diffuse light are set to 0.3 and 0.6 respectively. The strength of the specular reflection is set to 0. We apply the light uniformly across each triangular face of the 3D mesh (i.e. flat shading). All of the 42 views are then rendered at 200 dpi, also in OpenGL mode, and further resized to the resolution of 224×224.

4.4 Results of various object classes

Fig. 3 shows the distinction maps for a variety of 3D objects in our test dataset. In all the samples, the red regions are of the highest distinction and the blue regions are the least distinctive, and yellow and green regions are in the middle. One fundamental finding is that these distinction maps are highly consistent with human perception. For instance, the four feet of the quadrupeds are detected as the most distinctive features to the cow, the camel and the horse since we categorise them into 'Quadruped' due to the four feet rather than the head or the shape of the body which are not so distinctive. Similarly, the wheel of the car, the handle of the cup, the struts of the guitars and the facial region of the bust are important components for differing each class of objects within the ModelNet50 database. Another interesting finding is that similar local structures of objects in different classes could have very different distinction. For example, the handles of two cups are distinctive while the handle of one bottle is not distinctive at all. This is because while a handle is a very important structure for recognising cups, it is not a necessary element for a bottle.

4.5 Comparison with the distinction computed using handcrafted features

Fig. 4 compares our distinction maps with those shown in Figure 7 of [2] where all of the 15 person meshes are from the Princeton Shape Benchmark [50]. It can be seen that results shown in (a) and (b) are very different. In [2], mesh distinction was computed based on a popular hand-crafted feature, known as the Harmonic Shape Descriptor. Consequently, the elbow region is detected as the distinctive regions for persons while our method recognises the head and facial region as the most distinctive features and usually the hands as the second most distinctive. This is obviously more consistent with human perception since we can easily distinguish a person from other classes of objects by looking at these regions. As stated in [2], one limitation of their method is that "the distinctive regions may not correspond to semantic parts". In contrast, our method significantly better corresponds to semantic parts, which is typically desired as a perceptual measure. These results show that a sufficiently deep classification network in conjunction with an MRF is powerful for estimating the distinction of various 3D objects.

4.6 Results of single object class

To demonstrate that mesh distinction can reliably capture semantic regions in a consistent manner, we test it on the SHREC'15 dataset [51]. As shown in Fig. 5, mesh distinction behaves quite consistently although the persons have different gestures. The facial regions are always detected as the most distinctive features and the hands are usually the

Fig. 4. Distinction estimated through classification network and handcrafted features respectively. (a) Distinction computed by the proposed method based on classification network; (b) Distinction computed based on handcrafted features [2].

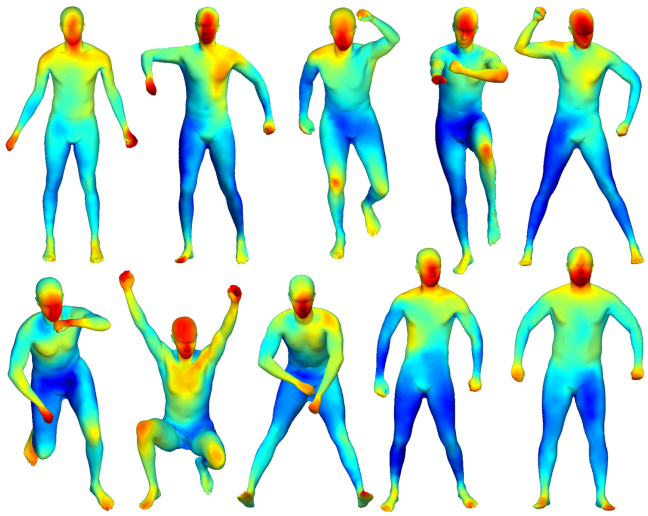


Fig. 5. Distinction of the same object class

second most distinctive. The consistency of such semantic regions is typically desired in applications such as shape matching and retrieval. It demonstrates that our method can distinguish the person class from other object classes even if a variety of non-rigid changes affect the people. We can see that the semantic parts captured by our method are not so different from those that draw our visual attention for recognition purpose.

4.7 Comparison with other perceptual measures

To further interpret the proposed distinction measure, we compare it with other perceptual measures by investigating how well mesh distinction corresponds to alternative perceptual measures across 20 object classes.

Mesh saliency A (MSa): [8] captured saliency through the spectral domain of the mesh. Global attributes are involved since the spectral processing is based on the entire mesh.

Mesh saliency B (MSb): [6] computed saliency by formulating several heuristics where global heuristics such as global shape extremities and global shape topology are considered.

Mesh saliency C (MSc): [35] also involves global cues where they used a local-to-global scheme to integrate both local and global saliency of features.

Mesh saliency D (MSd): [1] computed mesh saliency via Gaussian-weighted curvatures estimated in local neighbourhood at multiple scales and has no global consideration.

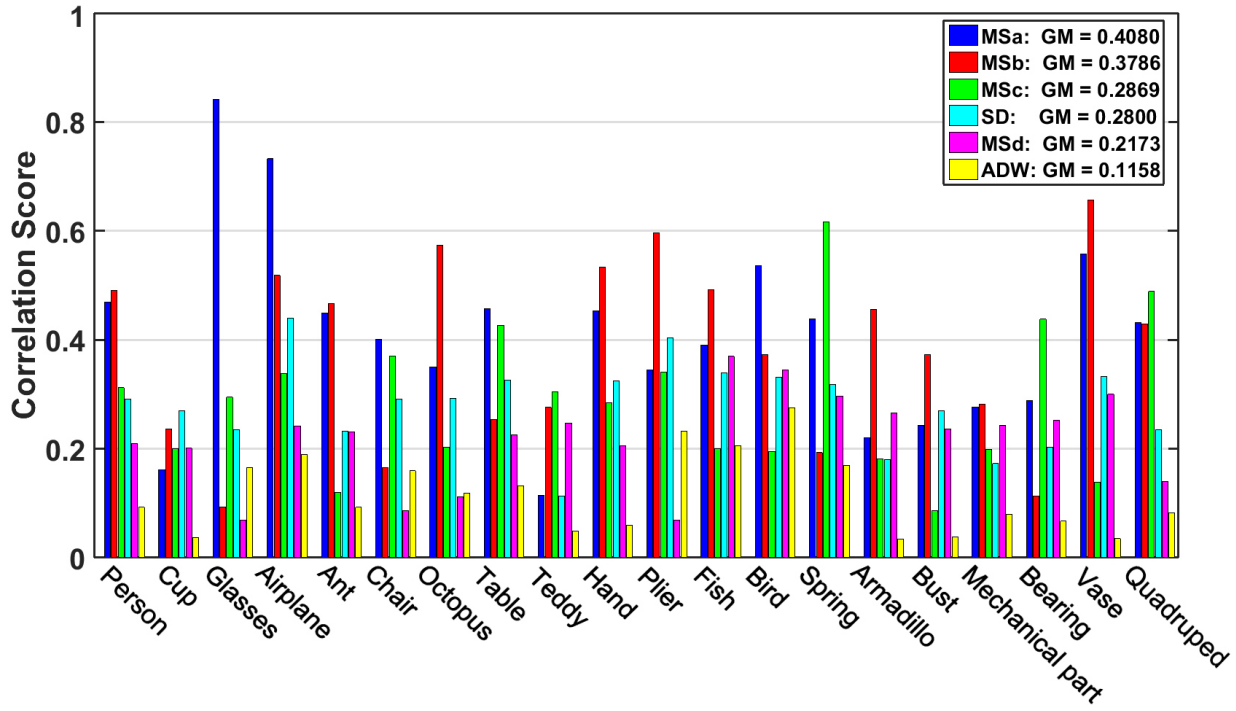


Fig. 6. Correlation score of the distinction measure with respect to other perceptual measures including [8] (MSa), [6] (MSb), [35] (MSc), [1] (MSd), [4] (SD) and [54] (ADW). Each bar denotes the mean value of the correlation scores of all test objects in a particular class. GM represents the global mean across all classes.

Schelling distribution (SD): [4] predicted the Schelling distribution by learning a regression tree where among the 13 attributes, only one (symmetry) is global.

Admissible diffusion wavelet (ADW): [54] computed saliency through admission diffusion wavelets constructed based only on a local operator.

We compared each of these measures to mesh distinction using the test models of 20 object classes[1]. Then, for a mesh containing M vertices, we calculate the absolute correlation score [55] to compare mesh distinction with other measures:

$$r = \left| \frac{1}{(M-1)\sigma_D \sigma_X} \sum_{m=1}^{M} (D(m) - \overline{D})(X(m) - \overline{X}) \right| \quad (12)$$

where $X(m)$ is one of the alternative measures computed at vertex m of a mesh and $D(m)$ is the distinction value of the vertex. $\overline{D}$ denotes the mean distinction over all vertices and $\overline{X}$ is the mean of X. σ_D and σ_X are the standard deviations of D and X respectively. Correlation scores close to 1 indicate a linear relationship and scores close to 0 indicate little or no association. To have an overview of the relationship between distinction and other perceptual measures, we calculate the mean of the correlation scores of the test models in the same object class and visualise the results across 20 classes in Fig. 6.

It can be seen in Fig. 6 that the correlation varies over object classes. For instance, distinction highly correlates to MSa [8] over the glasses but is quite independent of it over the Teddy and the cup. Also, distinction correlates well to MSb over the vase but weakly over the chair. When distinction strongly correlates to another perceptual measure over a specific object class, it typically means that the features that distinguish this class of objects from other classes are consistent with those captured by that perceptual measure in accordance with the particular cues it reasons about. Even so, we observed that in general, our distinction measure is more linearly correlated with the measures considering global cues. For example, in most classes, distinction has relatively strong correlation with MSa [8] but very weak correlation with ADW [54]. Note that in the MRF-CNN, each 2D view encodes global information of the input 3D object. And also the MRF globally combines information from multiple views. Thus global cues have a significant role in our method, which will be further demonstrated in the next section where we concern 3D scene containing multiple unconnected objects.

Overall, our distinction measure is fairly correlated to some perceptual measures over quite a few object classes, particularly the measures considering global attributes. This finding is against the point made in [2] that their distinction measure, computed based on handcrafted features, is extremely independent of other measures over all object classes. Probably, the reason is that the distinction computed based on a deep neural network is more powerful at capturing global perceptual attributes which are usually not so inconsistent compared with handcrafted local geometric attributes for a specific class of objects.

4.8 Applications and comparative evaluation using human-generated ground truth

We showcase two applications of mesh distinction computed by our method: best view selection and 3D inter-

1. Schelling distribution [4] is computed by learning from human-generated data acquired through a large scale user study which only covers 20 object classes. Therefore, the Schelling distribution of other object classes is not available.

TABLE 1
View Selection Errors of 8 view selections methods over 68 objects. VA: view area [52]; RVA: ratio of visible area [56]; SAE: surface area entropy [57]; SL: silhouette length [56]; SE: silhouette entropy [58]; CE: curvature entropy [58]; MS: mesh saliency [1]; MD: mesh distinction (our method)

MODELS	VA	RVA	SAE	SL	SE	CE	MS	MD
airplane	0.366	0.787	0.179	0.366	0.767	0.312	0.584	**0.167**
airplane2	**0.363**	0.656	0.622	0.637	0.534	**0.363**	0.590	0.617
airplane3	0.633	0.680	**0.320**	0.633	0.358	0.354	0.354	0.338
airplane_4	0.340	0.641	0.318	0.317	0.340	**0.230**	0.392	0.352
ant	0.716	0.692	0.467	0.601	0.496	0.357	**0.272**	0.692
armadillo	0.314	**0.176**	0.338	0.321	0.473	0.363	0.216	0.216
bicycle	0.695	0.273	0.354	0.677	**0.222**	0.423	0.591	0.293
bird_2	0.420	0.745	**0.228**	0.720	0.349	0.322	0.434	0.339
bird_3	0.673	0.647	0.647	0.373	0.644	**0.233**	0.583	0.644
bunny	0.788	0.649	0.673	0.212	0.564	0.454	0.873	**0.185**
bust	0.433	**0.122**	0.478	0.540	0.522	0.246	0.166	0.160
bust_2	0.552	0.883	**0.372**	0.552	0.480	0.724	0.493	0.493
cactus	0.202	0.627	0.269	**0.192**	0.570	0.617	0.320	0.254
camel	0.501	0.429	0.625	0.587	0.391	**0.335**	0.681	0.407
car	0.328	0.610	0.652	0.593	0.437	0.521	0.709	**0.235**
car2	0.668	0.502	0.568	0.631	0.653	0.610	**0.384**	0.584
car3	0.664	0.370	0.388	0.448	0.618	0.335	0.296	**0.203**
cat	0.855	0.564	0.365	**0.216**	0.328	0.450	0.365	0.618
chair	0.801	0.255	0.255	**0.221**	0.593	0.496	0.331	0.651
chair2	0.792	0.776	0.435	0.666	**0.208**	0.435	0.769	0.451
chair_4	0.769	0.296	0.366	0.366	0.620	**0.256**	0.751	0.726
chair_5	0.808	**0.249**	0.463	0.249	0.315	0.372	0.315	0.256
cow	0.286	0.612	0.427	0.316	**0.228**	0.321	0.398	**0.217**
cow_2	0.268	0.600	0.277	0.215	0.343	0.318	0.398	**0.096**
cup	0.618	0.352	0.368	0.562	0.368	0.535	**0.343**	0.657
desk_chair	0.186	0.175	0.175	0.451	0.730	0.670	**0.125**	0.185
dog	0.679	0.445	0.353	**0.321**	0.657	0.745	0.524	0.524
dog_2	0.548	0.562	0.562	0.647	0.425	**0.298**	0.562	0.562
dragon	0.645	0.483	0.611	0.425	0.443	**0.338**	0.573	0.509
duck	0.689	0.641	0.606	0.641	**0.403**	0.640	0.448	0.463
earthmover	0.651	0.380	0.594	0.652	0.693	0.416	0.643	**0.291**
feline	0.418	0.499	0.423	**0.356**	0.451	0.622	0.367	0.367
fish	0.464	**0.383**	0.383	0.441	0.568	0.413	0.500	0.625
flower	0.728	0.316	**0.267**	0.514	0.683	0.589	0.267	0.401
gargoyle	0.580	0.794	**0.214**	0.794	0.318	0.469	0.319	**0.214**
girl	**0.196**	0.699	0.226	0.233	0.212	0.598	0.305	0.623
glasses	0.450	0.698	0.637	**0.365**	0.410	0.396	0.698	0.662
guitar	**0.172**	0.222	0.222	0.245	0.582	0.698	0.503	0.250
hand2	0.669	0.361	0.361	0.612	0.290	0.710	0.218	**0.206**
hand_3	**0.271**	0.299	0.419	0.746	0.464	0.607	0.299	0.324
hand_4	0.550	0.454	0.456	0.519	0.539	**0.415**	0.603	0.556
head	0.635	0.213	0.592	0.271	0.355	0.588	**0.195**	0.213
helicopter	**0.348**	0.480	0.433	0.484	0.424	0.568	0.511	0.427
horse	**0.264**	0.384	0.478	0.579	0.692	0.355	0.323	0.595
horse_3	0.436	0.563	0.639	0.322	**0.311**	0.694	0.373	0.604
human	0.795	0.266	**0.114**	0.211	0.851	0.336	0.266	0.233
igea	0.506	0.877	0.426	0.516	**0.140**	0.496	0.490	0.489
lrover	0.590	0.644	0.166	0.349	0.607	0.546	0.696	**0.127**
maxplanck	**0.418**	0.422	0.422	0.497	0.453	0.463	0.422	0.365
nefertiti	0.670	0.641	0.476	0.697	0.565	0.526	0.524	**0.166**
octopus	0.529	0.474	**0.388**	0.526	0.562	0.646	0.420	0.508
piano	0.675	0.307	0.307	0.564	0.458	0.362	0.307	**0.214**
rabbit	0.364	0.315	**0.234**	0.550	0.562	0.483	0.255	0.289
rockerarm	0.463	0.534	0.519	**0.426**	0.569	0.605	0.459	0.459
santa	0.310	0.690	0.589	0.310	0.420	**0.261**	0.690	0.582
screwdriver	0.119	0.294	0.242	0.165	0.453	0.576	0.294	**0.090**
shape	0.704	0.336	0.420	0.670	0.724	0.705	0.322	**0.272**
shoe	0.614	0.487	**0.388**	0.590	0.449	0.551	0.415	0.310
shoe2	0.585	0.539	0.362	0.622	0.632	0.483	**0.313**	**0.313**
skull	0.579	0.769	0.310	0.427	0.682	0.393	0.377	**0.276**
table_2	0.636	0.338	0.425	0.365	0.508	**0.335**	0.636	0.353
teapot_2	0.447	0.367	0.367	0.410	0.360	0.619	0.367	**0.354**
teddy	0.191	0.244	0.286	**0.144**	0.448	0.551	0.244	0.191
utah_teapot	0.526	0.237	0.278	0.224	0.508	0.512	0.524	**0.209**
vase1	0.358	0.172	**0.125**	0.249	0.255	0.518	0.252	0.239
vase2	0.680	**0.282**	0.320	0.376	0.671	0.462	0.382	0.554
vase_1	0.492	0.471	0.466	0.488	0.610	0.642	0.471	**0.335**
wine_glass	0.494	0.180	0.209	**0.127**	0.340	0.339	0.180	0.457
AVERAGE	0.517	0.473	0.396	0.446	0.484	0.474	0.430	**0.380**

est point detection. In each case, we also quantitatively and comparatively evaluate distinction based on human-generated ground truth. We evaluate it through the two particular applications for three reasons. First, it is much easier to obtain human-generated ground truth for the two applications than distinction. Second, the ground truth of the two applications is more likely to be consistent and thus reliable. Third, publicly available benchmarks already exist. The results show that distinction is a very important attribute in both cases.

Best view selection. Perceptual measures [1], [6] have been applied to the selection of the best 2D view of a 3D object. Typically, the 'best' views are referred to as human preferred views [5], [6]. Thus researchers have released publicly available benchmark containing human-generated ground truth for evaluation purpose. Dutagaci *et al.* [52] provided a benchmark which involves a methodology, a quantitative measure and ground-truth best viewpoints generated by 26 people to evaluate view-selection algorithms.

To compute the best viewpoint in accordance with the distinction measure, we employ the simple scheme proposed in [1] which selects the viewpoint that maximises the sum of the saliency for the visible regions of the object. For a given viewpoint v, we find the best viewpoint as $v_b = \arg\max_v(\sum_{m \in S(v)} D(m))$ where $S(v)$ is the set of the vertices visible from v and $D(m)$ denotes the distinction of vertex m. Table 1 shows the performance of 8 competing methods over all of the 68 3D objects in the benchmark based on the View Selection Error (VSE) measure it provided. The VSE reflects the difference between the viewpoints selected by an algorithm and those selected by human subjects. It can be seen that our method based on mesh distinction (MD) has the smallest VSE (shown in bold font) over 19 objects and achieves the smallest average VSE. Perhaps it is not very safe to claim that our method definitely outperforms the other 7 methods considering that the benchmark is not large enough in terms of the number of human participants and the number of 3D objects it contains. However, we can conclude that distinction is definitely a very important attribute for selecting best views for a variety of objects.

3D interest point detection. Detection of interest points on a 3D surface is challenging since usually not just local attributes but some global attributes hard to compute are considered when people select them [3]. To extract a set of discrete distinct points from a continuous distinction distribution of a 3D surface mesh, we first remove the vertices with distinction values smaller than a global threshold; then we simply extract any vertex that either has a distinction value larger than a global threshold or is the local maximum.

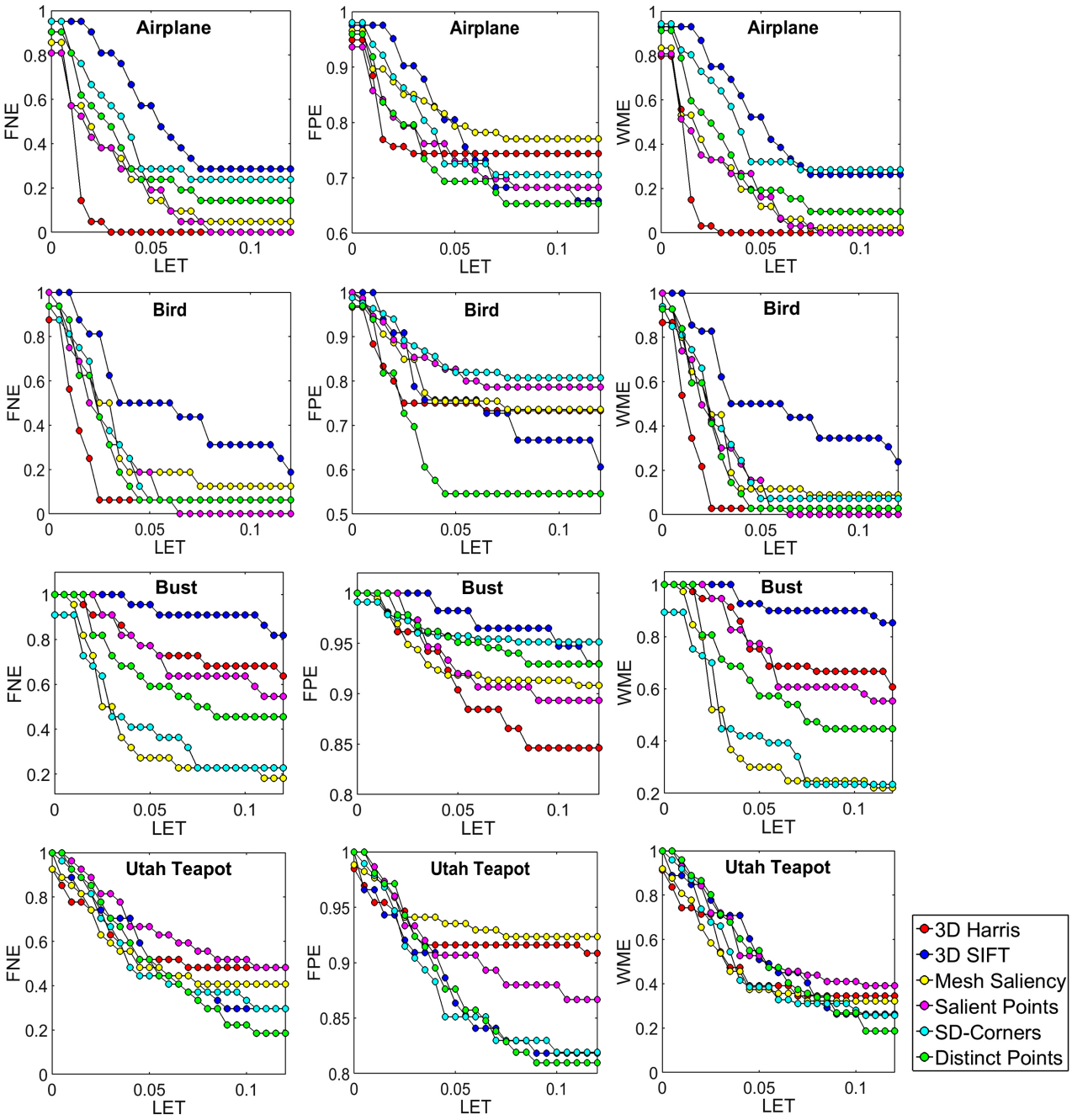


Fig. 7. FNE, WME, and FPE graphs of various algorithms of interest point detection for several test models

For a fair comparison, we test our method on a publicly available benchmark [3] which provides 3D interest points selected by human subjects as ground truth.

To measure how similar the detected points of interest are to those selected by human subjects, the benchmark proposes three metrics: false negative error (FNE), false positive error (FPE) and weighted miss error (WME). We compare our distinction-based method with the 5 competing methods suggested by the benchmark, namely 3D-Harris [59], 3D-SIFT [60], mesh saliency [1], salient points [30] and SD-corners [61]. Fig. 7 show the results through FNE, FPE and WME graphs with respect to the localisation error tolerance (LET). A vertex is considered to be 'correctly detected' as a point of interest if its geodesic distance to the closest ground truth is not larger than a specific LET value. FNE and FPE are defined in the obvious way. Normally, as more points of interest are captured, more false positives are detected while achieving a lower FNE. If a method tends to mark fewer points of interest, it results in a lower FPE, at the cost of a higher FNE. An algorithm gets a low WME if it manages to detect a point that is frequently voted by human subjects. Thus it measures the ability of a method to detect the most interesting points. In contrast, FNE and FPE treat all ground truth points of interest equally. Hence, an ideal method should keep FNE, FPE and WME all low. We can see that the distinct points derived from the distinction measure are at least comparable with the competing methods specifically designed for interest point detection. For the bird and the

TABLE 2
The effect of the MRF weighting parameter α

Values of α	Average of VSEs	Std of VSEs
0	0.422	0.193
0.05	0.402	0.183
0.1	0.380	0.173
0.15	0.391	0.178
0.2	0.401	0.181

teapot, it even outperforms all the competing methods. This means that quite a few distinct points are of human perceptual interest and many points of human interest can be used to distinguish the object from those of other object classes.

4.9 The effect of the MRF weighting parameter α

With the help of the ground truth data of view selection, we ran experiments to analyse the effect of the MRF weighting parameter α in Eq. (3). The results are shown in Table 2. $\alpha = 0$ can be regarded as an ablation analysis of the MRF model. It shows that by introducing the prior term U_p, both the average and the standard deviation of VSE over 68 meshes are significantly reduced. In this work, we empirically set $\alpha = 0.1$ which achieved the best performance. By scrutinising the VSE of each individual mesh, we found that most of them are not changed when replacing $\alpha = 0$ with $\alpha = 0.1$, which means that the distinctions of these meshes are not significantly changed. But some of them are reduced by more than 30%. A further investigation shows that if the VSE of a particular mesh changes significantly when replacing $\alpha = 0$ with $\alpha = 0.1$, usually some of the views of the mesh are incorrectly classified. In some cases, $P_{\mathcal{C}}(V_n)$ which represents the probability of a view V_n with respect to the correct class $\mathcal{C}$ is smaller than 0.3. Note that typically $P_{\mathcal{C}}(V)$ should be larger than 0.9. This means that the weight w_n (see Eq. (2)) corresponding to V_n will be incorrectly estimated if we rely only on the likelihood term U_l (Eq. (4)). However, the possibility that all of the neighbouring views of V_n are incorrectly classified is much smaller. Thus the prior term which penalises neighbouring views to take very different weights can mitigate such wrong estimation over w_n. Therefore, the prior term of the proposed MRF is essentially a robust strategy, which effectively improves distinction estimation when the classification is not reliable.

4.10 Comparison with Su's multi-view CNN [19]

Since our method integrates a multi-view CNN structure inspired by the work of Su *et al.* [19] with an MRF, it is interesting to see if such integration leads to significant improvement for mesh distinction. To enable a direct comparison, we replace the MRF component with the view-pooling layer proposed in [19] and all other components shown in Fig. 1 remain the same. Thus Su's multi-view CNN can be regarded as an ablated version of our method with the MRF component removed. The weighting parameters w_n in Eq. (2) are all set to 1 since they are not available with Su's method. We still use VSE as the metric to quantitatively indicate the performance. Fig. 8 shows the results of the comparison implemented on the same benchmark [52] used

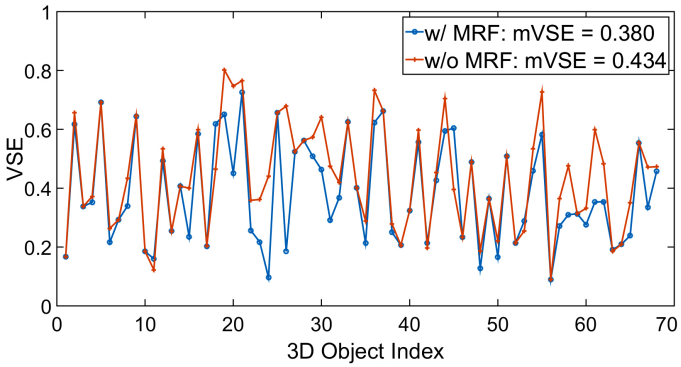


Fig. 8. The effect of the MRF component. We implement the view pooling layer proposed by *Su et al.* [19] as a substitute of the MRF in our method. mVSE denotes the mean VSE over 68 objects.

in Table 1 containing a variety of 68 objects. We can see that for most 3D objects, the combination of the multi-view CNN and the MRF leads to lower VSEs. On average, the proposed MRF significantly improves the performance over Su's multi-view CNN by more than 10% in terms of the VSE. In a multi-view CNN, a pooling operation such as Su's view pooling layer completely ignores the useful information about the relative locations of multiple viewpoints. But the proposed MRF is sensitive to it due to the prior term defined as the consistency between neighbouring views in Eq. (5).

5 Distinction of 3D Scenes

To the best of our knowledge, previous work on perceptual measures of 3D surface meshes only concerns a single object. In this work, for the first time, we extend the concept of mesh distinction from one single object to a 3D scene which contains multiple objects. Although perceptual measures of 3D scene might not be a new topic, to date researchers have only computed them based on depth images [22], [23], [24]. The disadvantage of depth-based perceptual measures of a 3D scene is that the resultant perceptual map is only valid for a particular view since depth is view-dependent and typically a large number of points which exist in the real scene have no perceptual values due to occlusion. In comparison, we aim at creating a per-vertex distinction map of a scene composed of multiple 3D meshes.

If we compute the distinction for each object separately, the challenge will be how to assign a weight to each of them representing the global importance of the object in the context of the scene. However, since there is no edge connecting the objects, it is very hard to estimate the relationship between them. The proposed method is based on a multi-view representation of the scene where each view encodes the information corresponding to the relationship between the objects. Such information is actually incorporated into the view-based pixel distinction maps in our method.

Figs. 9-14 show the results of applying our method to several 3D scenes (courtesy of 3D Warehouse [62]) where we directly applied the trained classification network for testing without any further training. Such a strategy is important due to the fact that acquisition of training data is even more difficult where extra effort for creating a scene from individual objects is needed. Our preliminary results show that it

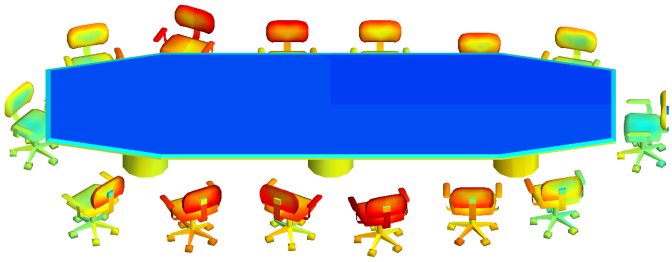


Fig. 9. Distinction of a conference room

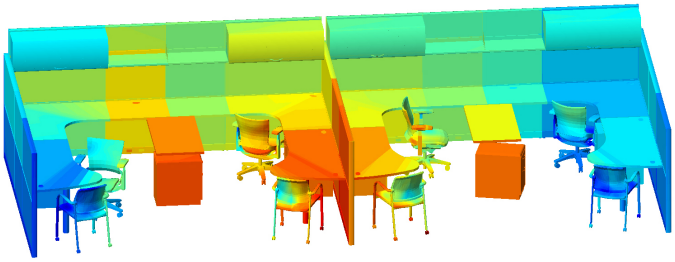


Fig. 10. Distinction of an office

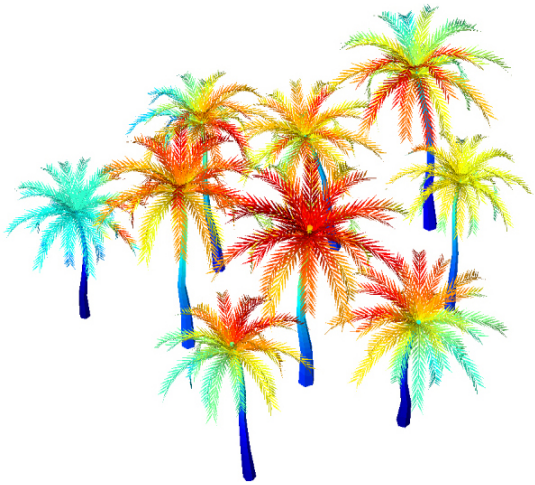


Fig. 11. Distinction of a scene containing several palm trees

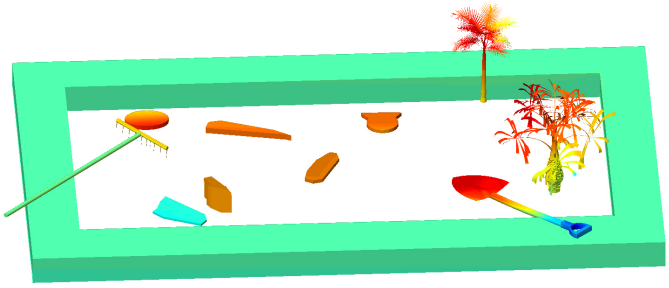


Fig. 12. Distinction of a garden

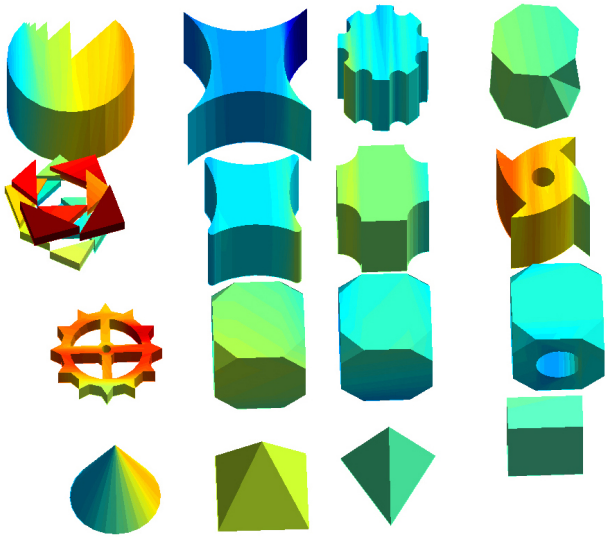


Fig. 13. Distinction of a scene containing multiple geometric shapes

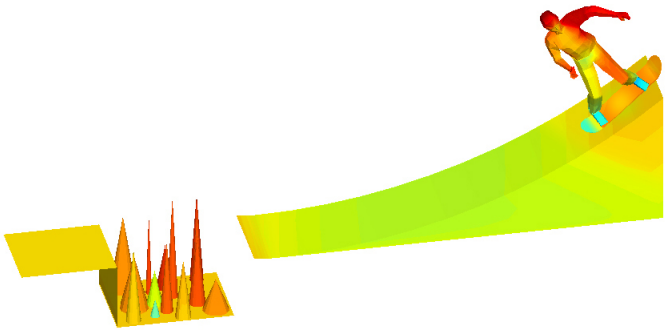


Fig. 14. Distinction of a skateboarding scene

is feasible to estimate the distinction of a scene through a classification network trained just on single objects. In the following, we summarise two interesting findings related to the behaviour of scene distinction we observed from these results.

Centre cue. We find that objects in the centre of the scene are likely to be distinctive. The chairs in the middle in the conference room scene (Fig. 9) are no different from the other chairs but are more distinctive than the chairs on the two ends of the scene. The office scene (Fig. 10) is more symmetric and the tables and the chairs in the middle are significantly more distinctive than those on the two sides. Fig. 11 further demonstrates the importance of centre cue where the palm trees in the central region are more distinctive than those at the corners. Note that people often frame the objects of interest near the centre of the image and the centre cue has been widely adopted as a top-down global cue in image saliency algorithms [27], [63], [64]. Centre cue can be regarded as a global cue of assigning weights to different objects and these results show that our distinction measure is consistent with it.

Discontinuity cue. However, local features such as the trees in Fig. 12, the gear wheel in Fig. 13, the cones and the person in Fig. 14 could still be distinctive even if they are at or close to the edges of the scenes. This is because compared to the objects in the central regions, they contain more discontinuities of depth and orientations. These findings echo the behaviour of image saliency [25], [65] where intensity and/or colour discontinuities usually correspond to salient regions.

These results highlight that our distinction measure

takes different cues into account and makes a good balance between them.

6 Conclusions and Future Work

We proposed a method for computing distinction, a perceptual measure which reflects the regional importance of a surface with regard to distinguishing the object from others of different classes. We found that most distinctive regions highlighted by our method usually correspond to semantic parts and are consistent with features of human interest. One weakness of the current work is that it is not completely invariant to object orientation as a view-based approach. One solution is to use a very large number of views to represent a 3D mesh since our MRF-based combination is invariant to viewpoint change. But this will significantly slow down the training and the deployment of the neural network. Under the current setting (42 views), the distinction of most test models can be computed within 1 minute on a computer with an Intel i7-4790 3.6GHz CPU and 32GB RAM without using any GPU acceleration. For some large input meshes such as the office scene containing 129K vertices, it took 149 seconds. Thus one future work is to develop a scheme as preprocessing for reliably generating canonical representations of 3D objects so that we merely need to use a very small number of canonical views to represent the object. Certainly, as a view-based method, it may generate incorrect distinction values for heavily occluded regions in meshes or scenes, especially in the case that a small number of views are used. So this scheme might also need to consider the degree of occlusion.

Another disadvantage of our method is that it is not trained end-to-end due to the introduction of the MRF. Thus an important direction for future work is to develop a deep architecture which can be trained end-to-end to further boost the performance.

In this work we showed some preliminary results for mesh-based 3D scene distinction, which introduces a new topic of interest. Its behaviour was summarised based only on the observations of a small number of scenes. Future work should extend such evaluation to larger datasets although collecting them could be difficult.

This work also reveals that features of 3D objects learned through a sufficiently deep CNN trained on classification datasets are transferable to other 3D object understanding tasks as long as proper heuristics related to the particular task are introduced to guide the feature selection/concatenation process. The lack of large-scale training dataset has become a bottleneck for many 3D object understanding tasks as such datasets are naturally much more difficult to create than their 2D counterparts. However, classification is an exception where the manual effort of creating 3D object classification training datasets and 2D image classification datasets is almost the same, and both are much easier than other object understanding tasks. Therefore, motivated by the performance of this work, another future work is to adapt the proposed method by considering new heuristics to other 3D object understanding tasks.

This work demonstrates that mesh distinction can be extended to 3D scenes. 3D scene understanding is an active topic mostly studied in a vision perspective and often implemented by using depth data. The mesh representation is usually more informative and less ambiguous than the depth representation of a 3D scene, so can it help to better understand the scene? We believe that mesh distinction and its potential invariants will become a good complement to the depth-based methods and even lead to a new, graphics perspective of 3D scene understanding in the future.

References

[1] C. H. Lee, A. Varshney, and D. W. Jacobs, "Mesh saliency," *ACM Trans. on Graph.(Proc. SIGGRAPH)*, vol. 24, no. 3, pp. 659–666, 2005.

[2] P. Shilane and T. Funkhouser, "Distinctive regions of 3D surfaces," *ACM Trans. on Graph.*, vol. 26, no. 2, p. 7, 2007.

[3] H. Dutagaci, C. Cheung, and A. Godil, "Evaluation of 3D interest point detection techniques via human-generated ground truth," *Vis. Compu.*, vol. 28, pp. 901–917, 2012.

[4] X. Chen, A. Saparov, B. Pang, and T. Funkhouser, "Schelling points on 3D surface meshes," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 31, no. 4, p. 29, 2012.

[5] A. Secord, J. Lu, A. Finkelstein, M. Singh, and A. Nealen, "Perceptual models of viewpoint preference," *ACM Trans. Graph.*, vol. 30, no. 5, p. 109, 2011.

[6] G. Leifman, E. Shtrom, and A. Tal, "Surface regions of interest for viewpoint selection," in *Proc. CVPR*, 2012, pp. 414–421.

[7] S. Wang, N. Li, S. Li, Z. Luo, Z. Su, and H. Qin, "Multi-scale mesh saliency based on low-rank and sparse analysis in shape feature space," *Comput. Aided Geom. Des.*, vol. 35, pp. 206–214, 2015.

[8] R. Song, Y. Liu, R. R. Martin, and P. L. Rosin, "Mesh saliency via spectral processing," *ACM Trans. on Graph.*, vol. 33, no. 1, 2014.

[9] W. Wang, H. Chao, J. Tong, Z. Yang, X. Tong, H. Li, X. Liu, and L. Liu, "Saliency-preserving slicing optimization for effective 3d printing," *Comput. Graph. Forum*, vol. 34, no. 6, 2015.

[10] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015.

[11] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. Venkatesh Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Proc. CVPR*, June 2016.

[12] K. Guo, D. Zou, and X. Chen, "3d mesh labeling via deep convolutional neural networks," *ACM Transactions on Graphics*, vol. 35, no. 1, p. 3, 2015.

[13] M. Lau, K. Dev, W. Shi, J. Dorsey, and H. Rushmeier, "Tactile mesh saliency," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 35, no. 4, 2016.

[14] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 38, no. 9, pp. 1901–1907, 2016.

[15] H. Zhang, X. Shang, W. Yang, H. Xu, H. Luan, and T.-S. Chua, "Online collaborative learning for open-vocabulary visual classifiers," in *Proc. CVPR*, 2016, pp. 2809–2817.

[16] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. CVPR (oral)*, 2017.

[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. NIPS*, 2014, pp. 3320–3328.

[18] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. ICML*, 2015.

[19] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proc. ICCV*, 2015, pp. 945–953.

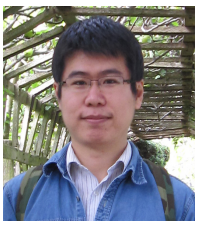
[20] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. Guibas, "Volumetric and multi-view cnns for object classification on 3d data," in *Proc. CVPR*, 2016, pp. 5648–5656.

[21] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3D shape segmentation with projective convolutional networks," in *Proc. CVPR*, vol. 1, no. 2, 2017, p. 8.

[22] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*. Springer, 2012, pp. 101–115.

[23] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. CVPR*, 2012, pp. 454–461.

[24] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, 2014, pp. 92–109.
[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 20, no. 11, pp. 1254–1259, 1998.
[26] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. CVPR*, 2007, pp. 1–8.
[27] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 34, no. 10, pp. 1915–1926, 2012.
[28] Y. Kim, A. Varshney, D. Jacobs, and F. Guimbretiere, "Mesh saliency and human eye fixations," *ACM Trans. Appl. Percept.*, vol. 7, no. 2, pp. 12:1–12:13, 2010.
[29] R. Gal and D. Cohen-Or, "Salient geometric features for partial shape matching and similarity," *ACM Trans. Graph.*, vol. 25, no. 1, pp. 130–150, 2006.
[30] U. Castellani, M. Cristani, S. Fantoni, and V. Murino, "Sparse points matching by combining 3d mesh saliency with statistical descriptors," in *Proc. Eurographics*, 2008, pp. 643–652.
[31] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.
[32] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.
[33] C. Koch and T. Poggio, "Predicting the visual world: silence is golden," *Nat. Neurosci.*, vol. 2, pp. 9–10, 1999.
[34] J. Wu, X. Shen, W. Zhu, and L. Liu, "Mesh saliency with global rarity," *Graph. Models*, vol. 46, pp. 264–274, 2013.
[35] R. Song, Y. Liu, R. Martin, and K. R. Echavarria, "Local-to-global mesh saliency," *Vis. Comput.*, vol. 34, no. 3, pp. 323–336, 2018.
[36] Z. Zhu, X. Wang, S. Bai, C. Yao, and X. Bai, "Deep learning representation using autoencoder for 3d shape retrieval," *Neurocomput.*, vol. 204, pp. 41–50, 2016.
[37] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proc. CVPR (oral)*, 2015, pp. 1912–1920.
[38] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. CVPR*, 2017.
[39] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in *Proc. NIPS*, 2016.
[40] L. Yi, H. Su, X. Guo, and L. J. Guibas, "Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation." in *Proc. CVPR*, 2017.
[41] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014.
[42] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
[43] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *PAMI*, vol. 26, no. 2, pp. 147–159, 2004.
[44] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *IJCV*, vol. 70, no. 1, pp. 41–54, 2006.
[45] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun, "Continuous Markov random fields for robust stereo estimation," in *Proc. ECCV*. Springer, 2012, pp. 45–58.
[46] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302, 1986.
[47] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C*. Cambridge Univ Press, 1982, vol. 2.
[48] T. A. Davis and W. W. Hager, "Row modifications of a sparse Cholesky factorization," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 3, pp. 621–639, 2005.
[49] H. Fu, D. Cohen-Or, G. Dror, and A. Sheffer, "Upright orientation of man-made objects," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 27, no. 3, p. 42, 2008.
[50] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Proceeds of Shape modeling applications*, 2004.
[51] D. Pickup et al., "SHREC'15 track: Shape retrieval of non-rigid 3D human models," in *Proceedings of the 7th Eurographics workshop on 3D Object Retrieval*, ser. EG 3DOR'15, 2015.
[52] H. Dutagaci, C. P. Cheung, and A. Godil, "A benchmark for best view selection of 3d objects," in *Proc. ACM workshop on 3DOR*, 2010, pp. 45–50.
[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*. IEEE, 2009, pp. 248–255.
[54] T. Hou and H. Qin, "Admissible diffusion wavelets and their applications in space-frequency processing," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 1, pp. 3–15, 2013.
[55] J. Pitman, *Probability*. Berlin, Germany: Springer-Verlag, 1993.
[56] O. Polonsky, G. Patané, S. Biasotti, C. Gotsman, and M. Spagnuolo, "What's in an image?" *The Visual Computer*, vol. 21, no. 8-10, pp. 840–847, 2005.
[57] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich, "Viewpoint selection using viewpoint entropy." in *VMV*, vol. 1, 2001, pp. 273–280.
[58] D. L. Page, A. F. Koschan, S. R. Sukumar, B. Roui-Abidi, and M. A. Abidi, "Shape analysis algorithm based on information theory," in *Proc. ICIP*, vol. 1, 2003, pp. I–229.
[59] I. Sipiran and B. Bustos, "Harris 3d: a robust extension of the Harris operator for interest point detection on 3d meshes," *The Visual Computer*, vol. 27, no. 11, pp. 963–976, 2011.
[60] A. Godil and A. Wagan, "Salient local 3d features for 3d shape retrieval," in *Proc. SPIE*, 2011.
[61] J. Novatnack and K. Nishino, "Scale-dependent 3d geometric features," in *Proc. ICCV*, 2007, pp. 1–8.
[62] [Online]. Available: https://3dwarehouse.sketchup.com/
[63] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. CVPR*, 2012, pp. 853–860.
[64] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. ICCV*, 2013, pp. 2976–2983.
[65] S. Lu, C. Tan, and J.-H. Lim, "Robust and efficient saliency modeling from image co-occurrence histograms," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 36, no. 1, pp. 195–201, 2014.



Dr Ran Song is a senior lecturer at the University of Brighton, UK since 2011. He received his Ph.D. in 2009 from the University of York, UK and his first degree in 2005 from Shandong University, China. He has published more than 30 papers in peer-reviewed international conferences proceedings and journals. His research interests lie in 3D shape analysis and 3D visual perception.

Prof Yonghuai Liu is a Professor at Edge Hill University, UK since 2018. Before his current post, he was a senior lecturer at Aberystwyth University, UK. He is currently associate editor and an editorial board member for a number of international journals, including Pattern Recognition Letters and Neurocomputing. He has published three books and more than 160 papers in international conference proceedings and journals. His primary research interests lie in 3D computer vision. He is a senior member of IEEE and Fellow of Higher Education Academy of United Kingdom.

Prof Paul L. Rosin is a Professor at the School of Computer Science and Informatics, Cardiff University, UK. Previous posts include lecturer at Brunel University London, UK, research scientist at the Institute for Remote Sensing Applications, Joint Research Centre, Ispra, Italy, and lecturer at Curtin University of Technology, Perth, Australia. His research interests include the representation, segmentation, and early image representations, low level image processing, machine vision approaches to remote sensing, methods for evaluation of approximation algorithms, etc., medical and biological image analysis, mesh processing, non-photorealistic rendering and the analysis of shape in art and architecture.